Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 145 (2016) 1314 – 1321

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Sustainable Design, Engineering and Construction

Latent class analysis for highway design and construction project categorization

Arthur L.C. Antoine[a]*, Keith R. Molenaar[b]

[a]*PhD Student, Dept. of Civil, Env. and Arch. Eng., University of Colorado Boulder, 428 UCB, Boulder, CO 80309-0428; email: arthur.antoine@colorado.edu*

[b]*Professor, Dept. of Civil, Env. and Arch. Eng., University of Colorado Boulder, 428 UCB, Boulder, CO 80309-0428; email: keith.molenaar@colorado.edu*

Abstract

Current classifications of highway design and construction projects are policy specific, ad hoc or based upon engineering judgement. This research defines an empirically-based highway project classification system through latent class analysis. Data from 291 projects completed between 2004 and 2015 by agencies across the United States serve as the basis for this analysis. The analysis explores project characteristic variables that include facility type, project type, highway type, project size in terms of cost and, project complexity. Latent class analysis provides an accurate and defensible project classification. This consistent classification of projects can aid researchers and practitioners in many applications such as enhancing the understanding of how agency decisions, like selection of a project delivery method, impact project performance.

Peer-review under responsibility of the organizing committee of ICSDEC 2016

Keywords: project descriptions, classification, project types, latent class analysis, highway construction

* Corresponding author. Tel.: 917-645-2095; fax: 303-492-7317.
E-mail address: Arthur.antoine@colorado.edu

Peer-review under responsibility of the organizing committee of ICSDEC 2016
doi:10.1016/j.proeng.2016.04.169

1. Introduction

The aim of this paper is to classify comprehensive descriptions of typical highway construction projects based on their work components and characteristics. Generally, researchers are focusing on ad hoc categories of projects but are neglecting the classification of descriptions within the pool of projects concerned. An empirical classification of projects can be complimentary to research efforts and can enhance results by showing how findings relate to specific projects. Such classification can also be beneficial to further analyses.

Currently, classifications of the descriptions of a projects' characteristics offered by researchers and highway engineers do not have an empirical basis when this could potentially benefit both academia and industry. General descriptions of construction projects exist but many of these descriptions are tied to policy or programmatic objectives, for example:

- "high priority projects" as referred to in Title 23 – Highways (23 USC 117) [1] for the high priority projects program; though, repealed July 6, 2012 by MAP-21 Section 1519;
- "complexity" definitions as provided by the National Cooperative Highway Research Program [2] to aid cost estimation efforts;
- "freight movement projects" as referred to in Title 23 – Highways (23 USC 150) [3].

The Federal Highway Administration (FHWA) defines a cost limit of $50M for "large" projects in policy but this is in specific relation to their value engineering federal-aid program [4]. Researchers are using some of these programmatic descriptions of projects in their work, e.g., Molenaar et al. (2007) [5], however, some researchers are still using subjective descriptions of projects, e.g., Debella and Ries (2006) [6] who refer to projects costing greater than $10M as "large", "complex" projects. Focusing on "large" projects with the assumption that these are very "complex" projects introduces ambiguities that can be avoided by the presentation of clear and distinct definitions of the descriptive terminology for highway construction projects. Regarding complexity, Gransberg (2013) [7] presents a graphical and quantitative means of assessing a project's complexity rated on five factors being technical, cost, context, financing and, schedule. This is a useful method for comparing projects on the basis of complexity however it does not present details of other significant project characteristics such as the facility type, project type or highway type.

Knowledge of project characteristics are critical in the practice of construction engineering and management. A notable statement from Gransberg et al. (2003) [8], reveals that there are critical project characteristics that can point to the use of a particular delivery method. Yet, statistically proven, empirical classifications of project characteristics are non-existent.

This research defines an empirically-based project classification of highway projects through latent class analysis (LCA). This consistent classification of projects captures the various conditions that are generic to highway construction projects and will aid researchers and agencies in many applications.

2. Data Collection

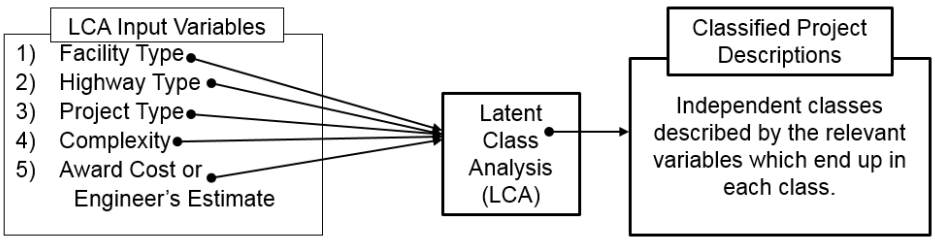

Fig. 1. project characteristics used as LCA input variables

Data from an ongoing FHWA national study [9] is used by the authors to define an empirically-based project classification of highway projects through LCA. The authors acquired information on project characteristics via a tested and well-structured questionnaire. The questionnaire was pre-approved by an FHWA's review panel and pilot tested with Departments of Transportation before distribution. Project characteristics are the input variables in LCA as shown in Fig 1. These project characteristics were defined through a review of highway agency scoping documents and these are factors known from the inception of a project. Twenty out of the 291 projects in the database have insufficient responses to the project characteristics to allow their use in LCA. Quality of the data was ensured both at the schema level and instance level by significant quality control techniques [10]. Double-checking responses with superior staff at the DOTs along with manual and low-level programming checks facilitated quality control in responses and data entry.

1. **Facility Type:** Respondents give the percentages of work components on the project in the categories of road, bridge, drainage, intelligent transportation systems (ITS) and other work, required to total 100%. Collectively, the responses for "other work" is re-labeled "Ancillary" work based on the provided qualitative explanations which include work such as landscaping, guardrail installation and signalization. The response, i.e. percentage, for each of these components of work are used as continuous variables in LCA.
2. **Project Type:** Respondents give the percentages of the project type within the descriptions of new construction/expansion, rehabilitation/reconstruction, resurfacing/renewal and others, required to total 100%. Collectively, the responses for "others" is re-labeled "Maintenance" work based on the provided qualitative explanations which include work descriptions such as maintenance, replacement and restoration. Observation of the raw data within the project type reveals that in the majority of cases, respondents classify projects as predominantly a single project type description. Thus, each of the project type descriptions are used as dichotomous nominal variables in LCA. Notably, use of these variables as categorical provides superior results than using them as continuous variables.
3. **Highway Type:** Respondents give the percentages of the highway type within the descriptions of rural interstate, urban interstate, rural primary, urban primary and rural secondary roads. As done with the project type, the highway type descriptions are dichotomous nominal variables in LCA because of the observation of the trend in responses; that is, respondents classify projects as predominantly a single highway type description.
4. **Complexity:** Each of the 3 complexity levels as defined in Table 1 is used as a dichotomous nominal variable in LCA. This provides superior results than using complexity as a 3 level polychotomous, categorical variable.
5. **Award Cost:** As a result of multiple trials to obtain appropriate bin sizes the award costs of the projects are arranged into categories of \$0-\$10M, \$10M-\$20M, \$20M-\$50M and, over \$50M. With these categories the award costs are input as 5 dichotomous nominal variables in LCA. In lieu of award cost data the engineer's estimate can be used in similar fashion.

Table 1. complexity definitions [2]

Most Complex (Major) Projects	Moderately Complex Projects	Non-complex (Minor) Projects
• New highways; major relocations • New interchanges • Capacity adding/major widening • Major reconstruction (4R; 3R with multi- phase traffic control) • Congestion management studies are required • Environmental Impact Statement or complex Environmental Assessment required	• 3R and 4R projects which do not add capacity • Minor roadway relocations • Non-complex bridge replacements with minor roadway approach work • Categorical Exclusion or non-complex Environmental Assessment required	• Maintenance betterment projects • Overlay projects, simple widening without right-of-way (or very minimum right-of-way take) little or no utility coordination • Non-complex enhancement projects without new bridges (e.g. bike trails) • Categorical Exclusion

"3R" = Resurfacing, Restoration, Rehabilitation

"4R" = New Construction/Reconstruction

3. Method of Analysis

Aside from having adequate data, consideration of an appropriate statistical method is imperative to be able to achieve the goal of classifying comprehensive descriptions of typical highway construction projects based on their work components and characteristics. A shortlist of potential statistical approaches for this purpose include factor analysis, cluster analysis and latent class analysis (LCA). Ultimately, LCA is selected because it provides the best means of meeting the goal of this research.

Cluster analysis, can illustrate how cases in a database congregate into separate groups, however, this approach has limitations and LCA provides significant advantages. Most notably, although it is also based on trends in responses, unlike LCA, cluster analysis is not a probabilistic model [11, 12]. In addition, LCA does not make any assumptions related to linearity, normal distribution or homogeneity [13, 14]. Thus, in the case of this study, cluster analysis cannot provide information such as the probability that given a project is rated as moderately complex, the project falls within a specific group. Alternatively, cluster analysis cannot provide information on the probability of a project being rated as moderately complex given that the project is within a specific group. LCA provides such conditional probabilities and in so doing, presents a convenient means of classifying observed and new cases; this is the main advantage of LCA over the clustering approach. The probabilistic methods of LCA results in LCA outperforming non-hierarchical cluster analysis methods such as k-means clustering and this contributes to LCA's usefulness by its ability to model quantitative data, especially mixed data types [15].

Factor analysis can reveal latent variables but this method is not ideally suited for use with mixed data types. Accurate factor analysis is best done on continuous variables, ideally from a normal distribution [16]. Nonetheless, with recent computational developments, many researchers are using factor analysis on ordinal data [17, 18]. Additionally, factor analysis is concerned with the structure of observed variables based on their correlation while LCA is more beneficial to this research because it can better present the structure of cases in a taxonomical form. Again, LCA is more advantageous because factor analysis does not present a convenient means of classifying observed and new cases.

LCA is a statistical method that can reveal unobservable groups of variables from observed/measured multivariate data based on the frequency of these variables and response patterns [11, 12, 19]. The groups of variables are referred to as latent classes which can then be used to classify cases from a database, as done in this study, and/or they can be used to confirm hypotheses about the resulting classification of cases. The latent classes or grouping of variables provides the means of classifying comprehensive descriptions of typical highway construction projects.

The basic latent class analysis model is given by the following equation which provides conditional probability also referred to as the posterior probability of class membership of measured variables [20]:

$$P(y_n \mid \theta) = \sum_{1}^{s} \pi_i P_j(y_n \mid \theta_j) \quad (1)$$

where y_n is the nth observation of the measured variables, S is the number of classes and π_i is the prior probability of membership in class j. P_j is the class specific probability of y_n given the class specific parameters θ_j. P_j will be probability mass functions when the measured variables are discrete and density functions when the measured variables are continuous. These conditional probabilities show how the latent classes differ and analysts can rename the latent classes based on the measured variables which fall within each class. The measured variables are assumed to be mutually independent within each latent class [14].With the posterior probability of class membership, maximum likelihood estimates can then be used to classify cases because LCA is a probabilistic model.

The Pearson chi square (X^2) and the likelihood ratio chi square (L^2) are criteria that can evaluate the goodness-of-fit of LCA models; analysts are also using the Akaike information criterion (AIC) and/or the Bayesian information criterion (BIC). With regard to the contingency table among variables, each criteria (X^2, L^2, AIC and BIC) compares the expected cell frequency count given by the resulting parameters of LCA with the actual cell frequency count from the data [19]. This is effectively a comparison of observed and expected response patterns. LCA results that produce expected cell frequency counts closest to actual cell frequency counts are acceptable. Consequently, LCA models having the lowest values within whichever criteria chosen are preferred. X^2 and L^2 can essentially test the hypothesis that the observed response frequency is equal to the expected frequency however, these chi square statistics have limitations. For instance, these chi square tests are useful when sample size is large and the number of input variables is small but they are invalid when there are too many sparse response patterns with low or zero frequencies. Analysts find the information criteria (AIC and BIC) advantageous for this reason. Supplemental to evaluating the goodness-of-fit, an option to assess the performance of LCA models is the estimated proportion of classification errors.

The steps in building the latent class model for this study include:

1. Identify and code the input variables.
2. Perform exploratory trials of LCA using different combinations of input variables.
3. Select the appropriate parsimonious model, having a suitably number of latent classes based on information criterion (BIC).
4. Use the posterior probabilities from LCA to assign each of the cases/projects in the database to one of the resulting latent classes.

The *"Data Collection"* section of this paper describes the variables used in LCA and how they are coded. Results from performing exploratory trials of LCA with different combinations of input variables enables the authors to identify and remove weak input variables; each trial containing multiple models having their own varying numbers of latent classes. The authors used Latent GOLD software version 5.0 by Statistical Innovations to perform LCA. In this study the variable of highway type is left out of the best LCA result because this is a confounding variable. Considering that highway projects are constructed in many different areas from rural to urban, many permutations of response patterns resulted in too numerous descriptive classes of projects and confounded results when the highway type is in the LCA.

4. Discussion

Ultimately, comparison of the evaluation criteria (BIC) by the authors reveals the best model with the ideal combination of input variables and an appropriate number of latent classes. The best model is a 3-class model in which the project characteristics were distinctly classified into separate latent classes as shown in Table 2.

Table 2. three (3) class model result from LCA

Variables / Classes	Descriptions			
	Complexity	**Award Cost**	**Facility Type**	**Project Type**
Class 1	Most Complex	Over $10M	Road & Drainage	New Const. & Resurf.
Class 2	Mod. Complex	$0 - $50M	Bridge	Rehab.
Class 3	Non-Complex	$0 - $10M	ITS & Ancillary	Maintenance

The 3-class model shown in Table 2 is the best result because the classification of project descriptions is easily interpretable, highly reflective of industry and maintains high average posterior probabilities (>0.90) for the classification of cases; associated BIC = 12501. Notably, the classification error rate (0.0053) is significantly better than the other models which used different combinations of input variables.

The 3-class model highly reflects what prevails in industry and provides an innovative and practical means of describing highway construction projects. Intuitive inspection reveals that there are strong logical ties between the raw data and the 3 latent classes which the authors appropriately titled.

Class 1 is titled as **"complex road construction"** projects. Class 1 includes variables that describe the most complex, new construction and resurfacing roadway projects costing over $10M. Unescapably and logically, drainage work is associated with the road projects. 56% of the 271 projects are classified as Class 1 projects.

Class 2 is titled as **"moderately complex bridge rehabilitation"** projects. Class 2 includes variables that describe the moderately complex, rehabilitation of bridges with costs ranging from $0 to $50M. 36% of the 271 projects are classified as Class 2 projects.

Class 3 is titled as **"non-complex ancillary and maintenance"** projects. Class 3 includes variables that describe the non-complex, ITS and ancillary, maintenance projects with costs ranging from $0 to $10M. 8% of the 271 projects are classified as Class 3 projects.

5. Conclusion

As researchers and highway agency officials continuously seek to improve design and construction performance the use of latent class analysis for variable reduction provides accurate and defensible classification of project descriptions. Aside from just variable reduction by providing a means of reasonably managing variables from a fairly large data set, LCA and the resulting consistent classification of projects can aid researchers and agencies in many applications such as:

i. To explore the relationship between projects' characteristics (e.g., size, complexity, facility type, etc.) and the resulting project performance (e.g., cost, schedule, intensity, etc.).
ii. Use by the construction industry as "preconditions" that prescribe the specific project delivery methods or other attributes of project delivery.
iii. For analysis of cost estimating to improve accuracy on the different projects.
iv. For assessment of project management efforts that are effective for different projects.

The resulting latent classes from LCA enable further analyses in other statistical approaches such as regression analysis or decision trees. To illustrate a Construction Engineering and Management research application, the latent classes can aid modeling project delivery in relation to project performance. For instance, given these new project descriptions researchers can examine which project delivery methods are more frequently used and the resulting project performance. Information from research such as this can help

agencies improve project delivery by enhancing efficiencies and mitigating against specific issues on different projects.

6. Limitations/Scope for Future Work

There exists project descriptions which are not revealed by the current latent class analysis as a result of the variables used in LCA and the sample size; for instance, "new construction, bridge projects" which were minimal (less than 11%) in the database of 291 projects and a classification for road projects with award costs less than $10 million. A more holistic data collection effort will reveal the full range of classifications of highway construction projects. Increasing sample size can also enhance the resulting descriptive classes by producing high conditional probabilities of membership for the input variables in the respective classes. This is simply the result of having an increased frequency of particular response patterns among the input variables. Additionally, increasing sample size can improve the classification of new and/or existing cases by producing clear probabilities of class membership.

Researchers should be wary that the resulting classification of projects is unique to the variables used in LCA. With more input variables, LCA will consequentially produce more descriptive classes of projects. However, these resulting descriptive classes may have less practical merit to industry or academia.

Acknowledgements

The authors extend their sincere gratitude to the numerous agency professionals who contributed to the nationwide study, to our partners at the FHWA for their assistance and advice, and to our research team partners, who provided invaluable assistance with quality control checks of the data.

References

[1] United States Code, Supplement 5, Title 23 - Highways. Title 23 - Highways, Chapter 1 - Federal-Aid Highways, Sec. 117 - High priority projects program, 2006 Edition, (23 USC 117).
[2] S. Anderson, K. Molenaar, C. Schexnayder, NCHRP Synthesis 574: Guidance for Cost Estimation and Management for Highway Projects During Planning, Programming, and Preconstruction, Transportation Research Board of the National Academies, Washington, D.C., 2007.
[3] National goals and performance management measures, Title 23 – Highways Section 150 (23 U.S. Code § 150), October 19, 2012.
[4] FHWA, Value Engineering Final Rule, 2014, pp. 52972- 52977, Vol. 79, No. 172. https://www.gpo.gov/fdsys/pkg/FR-2014-09-05/pdf/2014-21020.pdf Accessed December 30, 2015.
[5] K. Molenaar, S. Won, R. Smith, Expectations for accuracy in highway design-build cost estimates, ASCE/CIB, Construction Research Congress, Grand Bahama Island, May, 2007.
[6] D. Debella, R. Ries, Construction Delivery Systems: A Comparative Analysis of Their Performance within School Districts, J. Constr. Eng. Manage., 132(11), 2006, 1131–1138.
[7] D. Gransberg, J. Shane, K. Strong, C. Lopez del Puerto, Project complexity mapping in five dimensions for complex transportation projects, Journal of Management in Engineering, 2013, 29, 316-326.
[8] D. Gransberg, G. Badillo-Kwiatkowski, K. Molenaar, Project Delivery Comparison Using Performance Metrics, Association for the Advancement of Cost Engineering International (AACE International) Transactions, 2003, CS21-CS25.
[9] FHWA, Quantification of Cost, Benefits and Risk Associated with Alternative Contracting Methods and Accelerated Performance Specifications, Federal Highway Administration Project, Contract No. DTFH61-11-D-00009, 2013-2016.
[10] E. Rahm, H. Do, Data cleaning: problems and current approaches, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 23, Issue 4, 2000, pp. 3-13.
[11] J. Uebersax, LCA Frequently Asked Questions (FAQ), 2009, http://www.john-uebersax.com/stat/faq.htm, accessed July 11, 2015.
[12] Institute For Digital Research And Education (IDRE-UCLA), 2015, Mplus Data Analysis Examples, Latent Class Analysis, http://www.ats.ucla.edu/stat/mplus/dae/lca1.htm, accessed July 11, 2015.
[13] Statistics Solutions. Latent Class Analysis, 2015, https://www.statisticssolutions.com/latent-class-analysis/, accessed Aug. 11, 2015.

[14] J. Vermunt, J. Magidson, Latent class cluster analysis, in J. A. Hagenaars & A. L. McCutcheon (Eds.), Applied latent class models, 2002, pp. 89-106, Cambridge, UK: Cambridge University Press.
[15] J. Vermunt, J. Magidson, Latent class analysis, Encyclopedia of Social Science Research Methods, London: Sage, 2003.
[16] L. Fabrigar, D. Wegener, R. MacCallum, E. Strahan, Evaluating the use of exploratory factor analysis in psychological research, Psychological Methods, 1999, Vol.4. No. 3, 272-299.
[17] J. Meulman, A. Van Der Kooij, J. Heiser, Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data, in Kaplan, D., The SAGE Handbook of Quantitative Methodology for the Social Sciences. SAGE, 2004.
[18] IBM, Exploratory Factor Analysis with categorical variables, 2015, http://www-01.ibm.com/support/docview.wss?uid=swg21477550, accessed July 11, 2015.
[19] J. Hagenaars, A. McCutcheon, Applied Latent Class Analysis, Cambridge University Press, 2002.
[20] A. Eshghi, D. Haughton, P. Legrand, M. Skaletsky, S. Woolford, Identifying Groups: A Comparison of Methodologies, Journal of Data Science, 9(2011), 271-291.